# Assessing statistical significance in multivariable genome wide association analysis

Buzdugan, Laura ; Kalisch, Markus ; Navarro, Arcadi ; Schunk, Daniel ; Fehr, Ernst ; Bühlmann, Peter

Abstract: Motivation: Although Genome Wide Association Studies (GWAS) genotype a very large number of single nucleotide polymorphisms (SNPs), the data is often analyzed one SNP at a time. The low predictive power of single SNPs, coupled with the high significance threshold needed to correct for multiple testing, greatly decreases the power of GWAS. Results: We propose a procedure in which all the SNPs are analyzed in a multiple generalized linear model, and we show its use for extremely high-dimensional datasets. Our method yields p-values for assessing significance of single SNPs or groups of SNPs while controlling for all other SNPs and the family wise error rate (FWER). Thus, our method tests whether or not a SNP carries any additional information about the phenotype beyond that available by all the other SNPs. This rules out spurious correlations between phenotypes and SNPs that can arise from marginal methods because the "spuriously correlated" SNP merely happens to be correlated with the "truly causal" SNP. In addition, the method offers a data driven approach to identifying and refining groups of SNPs that jointly contain informative signals about the phenotype. We demonstrate the value of our method by applying it to the seven diseases analyzed by the WTCCC (The Wellcome Trust Case Control Consortium, 2007). We show, in particular, that our method is also capable of finding significant SNPs that were not identified in the original WTCCC study, but were replicated in other independent studies.

DOI: https://doi.org/10.1093/bioinformatics/btw128

# Supplementary Material

## S1 A causal interpretation

We consider here a linear structural equation model (Pearl, 2000):

$$X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} \gamma_{j,k} X_k + \varepsilon_j \ (j = 1, \ldots, p+1) \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_{p+1}$ are independent random variables with mean zero. Here, $X_{p+1} = Y$ denotes the random variable $Y$, and the structure of the model is given in terms of a directed acyclic graph $D$ which encodes direct causal effects: $\mathrm{pa}(j)$ are the parents of node $j$ in the graph $D$ and thus, $\mathrm{pa}(j)$ corresponds to the variables which have a direct causal effect to $X_j$ (arrows which point into node $j$).

**Proposition S1.1.** *Assume that the variables $X_1, \ldots, X_p, Y$ come from a linear structural equation model as in (1) where $Y$ is childless (i.e. all directed edges connected to $Y$ must point into $Y$), and consider the linear model representation as in eq. (1) in Section 2.1. Then: if $\beta_j \neq 0$ there is a directed edge $X_j \to Y$, and hence a direct causal effect from $X_j$ to $Y$.*

*Proof of Proposition S1.1.* We order the variables $X_1, \ldots, X_p, Y$ such that the structural equation model in (1) can be written as

$$X = BX + \varepsilon, \quad X = (X_1, \ldots, X_p, X_{p+1} = Y)^T$$

with a strictly lower triangular matrix $B$ and $\varepsilon_j$ being independent of $X_1, \ldots, X_{j-1}$ $(j = 1, \ldots p+1)$ (Loh and Bühlmann, 2014, cf.). The $j$th row of $B$ encodes the coefficients of the $j$th structural equation: $B_{j,k} = \gamma_{j,k}$ for $k \in \mathrm{pa}(j)$ and $B_{jk} = 0$ otherwise. Even if we don't know the order of the variables, but since $Y$ is childless, we can place it as the last variable in the order and write:

$$Y = \sum_{k=1}^{p} B_{p+1,k} X_k + \varepsilon_{p+1},$$

where $\varepsilon_{p+1}$ is independent of $X_1, \ldots, X_p$. This is an ordinary linear regression model for $Y$ versus all other variables $X_1, \ldots, X_p$ as in eq. (1) in Section 2.1 with $\beta_j = B_{p+1,j}$.

Therefore, if the regression coefficient $\beta_j = B_{p+1,j} \neq 0$, and since the non-zero values of $B$ encode for directed edges in the structural equation model, we must have that there is directed edge $X_j \to Y$. □

## S2 P-values for the hierarchical inference procedure

The hierarchical inference procedure described in Section 2.4 has been developed in Mandozzi and Bühlmann (2015), providing mathematical guarantees of familywise error control for generalized linear models. The main assumptions can be summarized as follows.

**1: Sparsity:** The generalized linear regression of the response $Y$ against the regressors $X_j$ $(j = 1, \ldots, p)$ is sparse. In particular, the active set

$$S = \{j;\ \beta_j \neq 0\}$$

has cardinality $s = |S|$ of smaller order of magnitude than $\sqrt{\log(p)/n}$.

**2: Beta-min:** The non-zero coefficients $\beta_j$ are sufficiently large. In particular, $\min_{j \in S} |\beta_j|$ is of larger order of magnitude than $s\sqrt{\log(p)/n}$.

**3: Identifiability:** The $n \times p$ design matrix, which contains all the regressors, arises as i.i.d. realizations of rows, each of them from the (same) sub-exponential [1] $p$-dimensional distribution with covariance matrix whose minimal eigenvalue is bounded away from zero

The proposed modification here for logistic regression goes along the lines as in Mandozzi and Bühlmann (2015), but requiring to consider the properties of logistic Lasso in the logistic regression model in eq. (2) in Section 2.1, see for example Bühlmann and van de Geer (2011). Furthermore, for step 3 of the procedure described in Section 2.4.2, we use the likelihood ratio test comparing the fit of the full and a reduced model (instead of a t- or F-test for linear models).

## S3 $R^2$ Computation

We use a scaled version of the generalized $R^2$, proposed by Nagelkerke (1991). The $R^2$ of a cluster or group $G$, in the $b^{th}$ split (step 1 in the procedure described in Section 2.4.2) is defined as follows:

$$R^2_{G,b} = \frac{1 - \left(\frac{L_0}{L(\hat{\theta}_{G,b})}\right)^{2/n}}{1 - (L_0)^{2/n}},$$

where $L(\hat{\theta}_{G,b})$ denotes the likelihood of the fitted model, and $L_0$ the likelihood of the null model. The SNPs that enter in the computation of $L(\hat{\theta}_{G,b})$ are a subset of the ones selected by the Lasso, in step 2 of the procedure described in Section 2.4.2. The aggregated $R^2$ of a cluster or group $G$ is then:

$$R^2_G = \frac{1}{B} \sum_{b=1}^{B} R^2_{G,b}$$

[1] For SNPs with discrete values, the assumption of sub-exponentiality is always fulfilled.

# S4 Results

## S4.1 Data preprocessing

Before running the analysis, we excluded the samples and SNPs provided in the WTCCC exclusion lists. We used PLINK (Purcell *et al.*, 2007) to further remove:

1. Samples with missing data above 10%.
2. SNPs with missing data above 10%.
3. SNPs with Minor Allele Frequency (MAF) $< 0.01$.
4. SNPs with extreme departure from the Hardy-Weinberg equilibrium ($P < 10^{-4}$).
5. SNPs located on the X chromosome.

After these exclusions, the data set consists of approximately $380'000$ SNPs. Due to the fact that multiple regression models cannot be fitted if the predictors have missing values, we used SHAPEIT (Delaneau *et al.*, 2013) to impute the missing SNP values. SHAPEIT phases the chromosomes using the NCBI Build 38 coordinate system of the Human Genome, and it automatically imputes the missing values of the genotyped SNPs.

We performed average linkage hierarchical clustering for the SNPs in each chromosome, ending up with one hierarchy per chromosome. These hierarchies were then joined into a final one which contained all the SNPs we used in our study.

## S4.2 Coronary artery disease

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2d}$ | Hits[e] |
|---|---|---|---|---|
| 17564 (58 %) | 1 | 0.032 | 0.022 | 5 out of 12 |
| 12965 (42 %) | 1 | 0.019 | 0.018 | 7 out of 12 |
| 11393 (35 %) | 2 | 0.049 | 0.016 | 4 out of 6 |
| 9563 (36 %) | 3 | 0.005 | 0.017 | 1 out of 15 |
| 24827 (100 %) | 4 | 0.014 | 0.026 | 4 out of 4 |
| 7113 (28 %) | 5 | 0.045 | 0.015 | 13 out of 14 |
| 21919 (91 %) | 6 | 0.023 | 0.025 | 2 out of 2 |
| 11458 (56 %) | 7 | 0.024 | 0.017 | 3 out of 3 |
| 21606 (100 %) | 8 | 0.039 | 0.024 | 6 out of 6 |
| 15116 (100 %) | 13 | 0.028 | 0.020 | 3 out of 3 |
| 6917 (63 %) | 15 | 0.047 | 0.013 | 2 out of 3 |

Table S1: **List of large significant groups of SNPs selected by our method for coronary artery disease.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs in individual chromosomes.

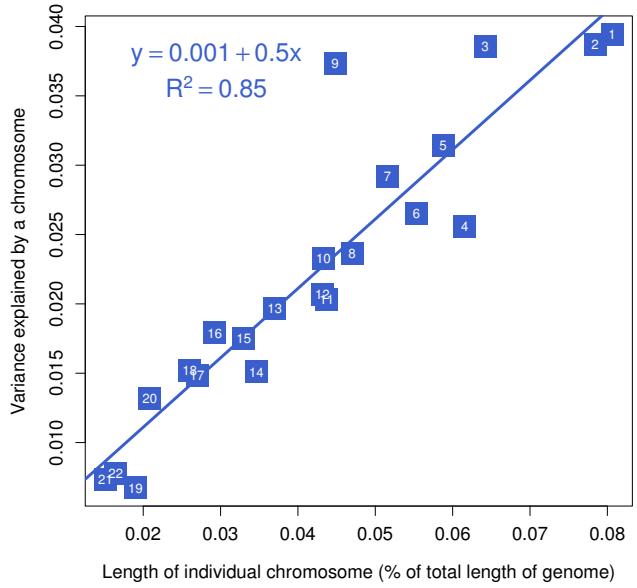


Figure S1: **Variance in coronary artery disease that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.49.

## S4.3 Crohn's disease

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2}$[d] | Hits[e] |
|---|---|---|---|---|
| 8201 (27 %) | 1 | 0.011 | 0.015 | 8 out of 35 |
| 5411 (21 %) | 3 | 0.030 | 0.013 | 23 out of 31 |
| 18313 (74 %) | 4 | 0.0009 | 0.025 | 5 out of 7 |
| 6490 (25 %) | 5 | 0.011 | 0.012 | 21 out of 41 |
| 6940 (29 %) | 6 | 0.023 | 0.010 | 4 out of 11 |
| 17262 (71 %) | 6 | 0.010 | 0.021 | 7 out of 11 |
| 20636 (100 %) | 7 | 0.009 | 0.026 | 12 out of 12 |
| 15434 (71 %) | 8 | 0.003 | 0.017 | 6 out of 7 |
| 18238 (100 %) | 9 | 0.040 | 0.018 | 3 out of 3 |
| 7902 (64 %) | 14 | 0.044 | 0.012 | 5 out of 8 |
| 8703 (100 %) | 17 | 0.010 | 0.015 | 7 out of 7 |

Table S2: **List of large significant groups of SNPs selected by our method for Crohn's disease.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

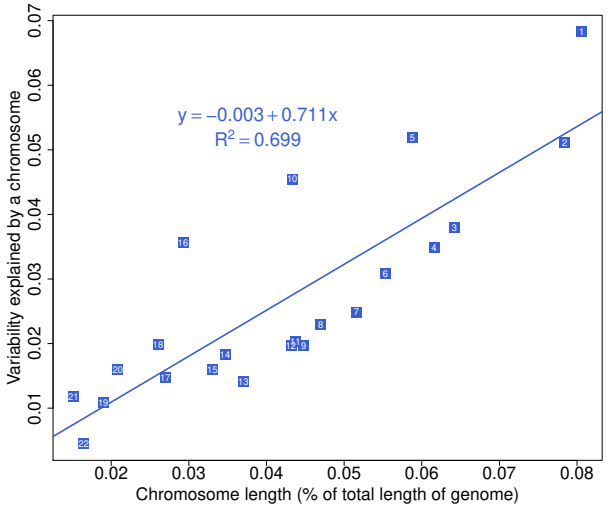


Figure S2: **Variance in Crohn's disease that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.55.

## S4.4 Rheumatoid arthritis

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2d}$ | Hits[e] |
|---|---|---|---|---|
| 13498 (42 %) | 2 | 0.014 | 0.018 | 1 out of 2 |
| 21287 (86 %) | 4 | 0.009 | 0.026 | 4 out of 4 |
| 21832 (86 %) | 5 | 0.003 | 0.027 | 4 out of 4 |
| 11669 (57 %) | 7 | 0.001 | 0.023 | 8 out of 9 |
| 5073 (28 %) | 9 | 0.030 | 0.011 | 2 out of 2 |
| 22512 (100 %) | 10 | 0.010 | 0.029 | 10 out of 10 |
| 14951 (73 %) | 11 | 0.026 | 0.018 | 6 out of 6 |
| 18315 (93 %) | 12 | 0.038 | 0.022 | 4 out of 4 |
| 9687 (82 %) | 16 | 0.015 | 0.013 | 3 out of 3 |

Table S3: **List of large significant groups of SNPs selected by our method for rheumatoid arthritis.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

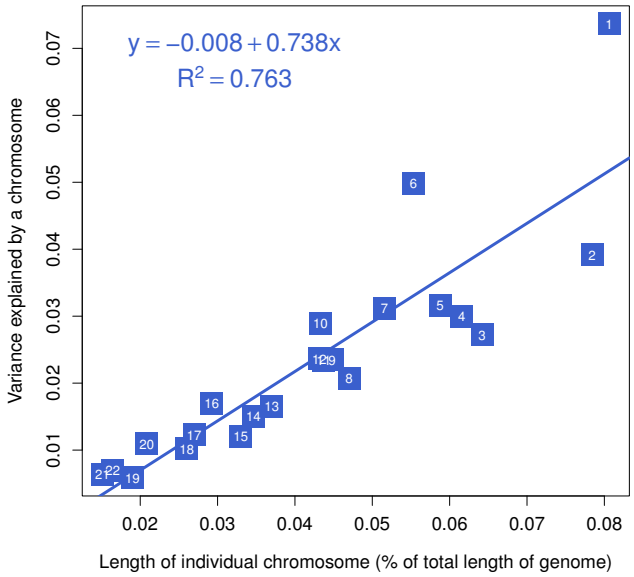


Figure S3: **Variance in rheumatoid arthritis that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.5.

## S4.5 Type 1 diabetes

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2d}$ | Hits[e] |
|---|---|---|---|---|
| 3057 (10 %) | 1 | 0.032 | 0.009 | 10 out of 42 |
| 12095 (38 %) | 2 | 0.034 | 0.020 | 8 out of 16 |
| 10366 (32 %) | 2 | 0.032 | 0.014 | 6 out of 16 |
| 15209 (57 %) | 3 | 0.006 | 0.021 | 9 out of 11 |
| 6699 (27 %) | 4 | 0.037 | 0.010 | 0 out of 7 |
| 7559 (30 %) | 5 | 0.012 | 0.010 | 4 out of 14 |
| 11473 (45 %) | 5 | 0.0003 | 0.017 | 10 out of 14 |
| 21450 (89 %) | 6 | 0.008 | 0.022 | 2 out of 101 |
| 116 (0.5 %) | 6 | 0.003 | 0.017 | 86 out of 101 |
| 20640 (100 %) | 7 | 0.007 | 0.022 | 3 out of 3 |
| 14694 (68 %) | 8 | 0.011 | 0.014 | 0 out of 2 |
| 11104 (61 %) | 9 | 0.025 | 0.015 | 4 out of 4 |
| 7418 (33 %) | 10 | 0.033 | 0.013 | 5 out of 7 |
| 18617 (91 %) | 11 | 0.014 | 0.019 | 2 out of 3 |
| 8441 (56 %) | 13 | 0.033 | 0.009 | 7 out of 8 |
| 11875 (97 %) | 14 | 0.029 | 0.016 | 0 out of 1 |
| 11051 (100 %) | 15 | 0.003 | 0.019 | 3 out of 3 |
| 8696 (100 %) | 17 | 0.010 | 0.014 | 1 out of 1 |
| 10948 (95 %) | 18 | 0.021 | 0.015 | 3 out of 3 |

Table S4: **List of large significant groups of SNPs selected by our method for type 1 diabetes.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

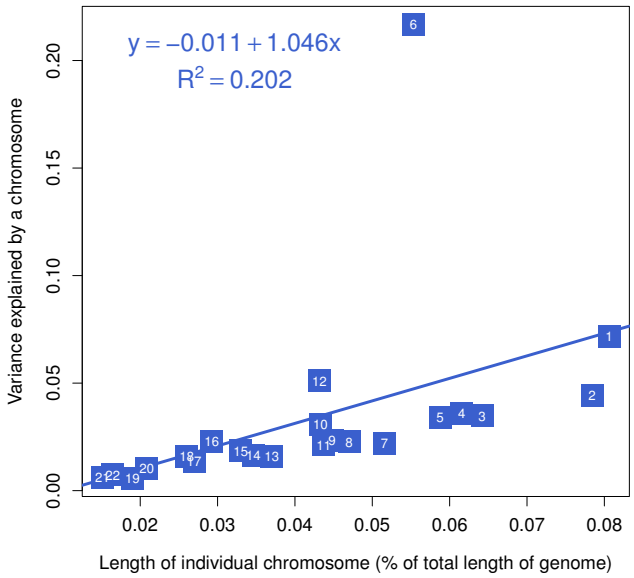


Figure S4: **Variance in type 1 diabetes that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.65.

## S4.6 Type 2 diabetes

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2d}$ | Hits[e] |
|---|---|---|---|---|
| 11446 (37 %) | 1 | 0.036 | 0.017 | 1 out of 5 |
| 8391 (26 %) | 2 | 0.018 | 0.013 | 13 out of 17 |
| 20050 (76 %) | 3 | 0.015 | 0.028 | 6 out of 7 |
| 10207 (41 %) | 4 | 0.045 | 0.016 | 2 out of 6 |
| 25440 (100 %) | 5 | 0.032 | 0.029 | 0 out of 0 |
| 21968 (91 %) | 6 | 0.011 | 0.026 | 7 out of 15 |
| 20645 (100 %) | 7 | 0.002 | 0.027 | 1 out of 1 |
| 20780 (96 %) | 8 | 0.038 | 0.022 | 5 out of 5 |
| 14962 (82 %) | 9 | 0.008 | 0.020 | 5 out of 5 |
| 4960 (25 %) | 12 | 0.039 | 0.011 | 15 out of 22 |
| 12261 (100 %) | 14 | 0.043 | 0.014 | 3 out of 3 |

Table S5: **List of large significant groups of SNPs selected by our method for type 2 diabetes.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

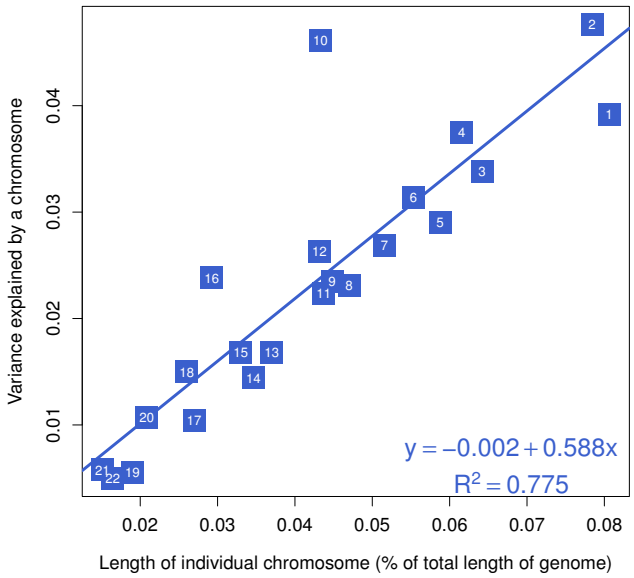


Figure S5: **Variance in type 2 diabetes that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.5.

## S4.7 Hypertension

| Size of significant SNP group[a] | Chr[b] | P-value[c] | $R^{2d}$ | Hits[e] |
|---|---|---|---|---|
| 19821 (65 %) | 1 | 0.010 | 0.023 | 1 out of 12 |
| 15595 (48 %) | 2 | 0.042 | 0.017 | 5 out of 9 |
| 26495 (100 %) | 3 | 0.004 | 0.035 | 10 out of 10 |
| 20019 (81 %) | 4 | 0.024 | 0.023 | 6 out of 7 |
| 25454 (100 %) | 5 | 0.046 | 0.027 | 2 out of 2 |
| 16965 (70 %) | 6 | 0.013 | 0.024 | 7 out of 9 |
| 20634 (100 %) | 7 | 0.021 | 0.025 | 7 out of 7 |
| 5436 (24 %) | 10 | 0.038 | 0.012 | 21 out of 21 |
| 20501 (100 %) | 11 | 0.002 | 0.024 | 5 out of 5 |
| 10428 (53 %) | 12 | 0.010 | 0.016 | 10 out of 14 |
| 15103 (100 %) | 13 | 0.007 | 0.022 | 5 out of 8 |
| 11052 (100 %) | 15 | 0.010 | 0.017 | 2 out of 2 |
| 9839 (100 %) | 20 | 0.037 | 0.018 | 4 out of 4 |

Table S6: **List of large significant groups of SNPs selected by our method for hypertension.**
[a] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.
[b] The chromosome to which the SNPs in the group belong.
[c] The p-value of the group of SNPs, adjusted for multiple testing (controlling the FWER).
[d] The variance explained by the group of SNPs.
[e] We counted the number of SNPs with p-values $< 5 * 10^{-4}$ identified using PLINK (Purcell *et al.*, 2007). We looked at how many of those SNPs are present in the groups selected by our method. The numbers refer to the SNPs on individual chromosomes.

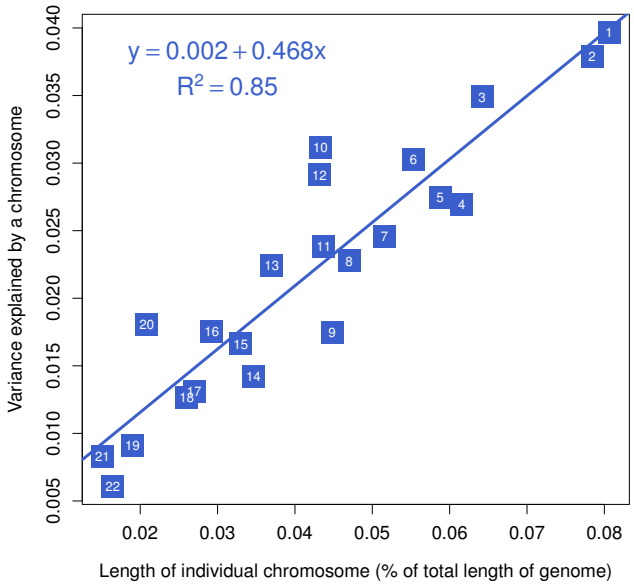


Figure S6: **Variance in hypertension that is explained by individual chromosomes.** The variance on the vertical axis is given by the $R^2$ value of all the selected SNPs on a chromosome, as described in the Supplementary Material Section 3. The total variance explained by all the selected SNPs on all the chromosomes is 0.48.

## S4.8 Stability of results

Due to the sample sample splitting step , our method has an element of randomness. The large number of sample splits are chosen to remove this variability. However, to show that the results are stable, we reran the analysis five times for Crohn's disease. The small and large significant clusters are detailed in Tables S7 and S8. As one can see, the changes between the different runs are very minor, especially for the small clusters.

Table S7: **Assessing the stability of the algorithm, by rerunning the analysis on the Crohn's disease data. The list of small clusters identified by our method for each of the five runs.**
[a] The chromosome to which the SNPs in the group belong.
[b] The list of small clusters or individual SNPs selected by our method in the current run.

| Chr[a] | Run 1[b] | Run 2[b] | Run 3[b] | Run 4[b] | Run 5[b] |
|---|---|---|---|---|---|
| 1 | rs12141431, rs11209033, rs12119179, rs11805303, rs2201841 | rs12141431, rs11209033, rs12119179, rs11805303, rs2201841 | rs6660226, rs11209039, rs12141431, rs11209033, rs12119179, rs11805303, rs2201841, rs10489629, rs10489628, rs4655684, rs4655679, rs10789224, rs17375018, rs11209018, rs6664119, rs7539795, rs6588245 | rs6660226, rs11209039, rs12141431, rs11209033, rs12119179, rs11805303, rs2201841 | rs6660226, rs11209039, rs12141431, rs11209033, rs12119179, rs11805303, rs2201841, rs10489629, rs10489628, rs4655684, rs4655679, rs10789224, rs17375018, rs11209018, rs6664119, rs7539795, rs6588245 |
| 2 | rs10210302 | rs10210302 | rs10210302 | rs10210302 | rs10210302 |
| Continued on next page | | | | | |

**Table S7 – continued from previous page**

| Chr[a] | Run 1[b] | Run 2[b] | Run 3[b] | Run 4[b] | Run 5[b] |
|---|---|---|---|---|---|
| 5 | rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934 | rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934 | rs11954639, rs6889990, rs6878963, rs16870155, rs6882351, rs10055946, rs7718129, rs1876143, rs7718309, rs16870170, rs13181692, rs4434422, rs2135330, rs10941516, rs16900114, rs10473203, rs10055860, rs11750156, rs1122433, rs1505992, rs4957317, rs4957313, rs1553576, rs1553577, rs6896604, rs6866402, rs2329353, rs10074991, rs13361707, rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934, rs17226632, rs6883686, rs6885315, rs9292776, rs6897022 , rs12658567 | rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934 | rs6871834, rs4957295, rs11957215, rs10213846, rs4957297, rs4957300, rs9292777, rs10512734, rs16869934 |
| Continued on next page | | | | | |

**Table S7 – continued from previous page**

| Chr[a] | Run 1[b] | Run 2[b] | Run 3[b] | Run 4[b] | Run 5[b] |
|---|---|---|---|---|---|
| 10 | rs10883371 | rs10883371, rs10883367, rs10883365, rs1548962 | rs10883367, rs10883365, rs1548962 | rs6584283, rs7095491, rs10883371, rs10883367, rs10883365, rs1548962 | rs10883371, rs10883367, rs10883365, rs1548962 |
| 10 | rs10761659 | rs10761659 | rs10761659 | rs10761659 | rs10761659 |
| 16 | rs2076756 | rs2076756 | rs2076756 | rs2076756 | rs2076756 |
| 18 | rs2542151 | rs16939895, rs7234029, rs2542151, rs2847297 | rs2542151 | rs2542151 | rs2542151 |

| Chr[a] | Run 1[b] | Run 2[b] | Run 3[b] | Run 4[b] | Run 5[b] |
|---|---|---|---|---|---|
| 1 | 8201 (27 %) | 3621 (12 %) | 8201 (27 %) | 3621 (12 %) | 3621 (12 %) |
| 3 | 5411 (21 %) | 18161 (69 %) | 18161 (69 %) | 18161 (69 %) | 18161 (69 %) |
| 4 | 18313 (74 %) | 18313 (74 %) | 18313 (74 %) | 18313 (74 %) | 18313 (74 %) |
| 5 | 6490 (25 %) | 17289 (68 %) | | 17289 (68 %) | 17289 (68 %) |
| 6 | 6940 (29 %) | 6940 (29 %) | | | 6940 (29 %) |
| 6 | 17262 (71 %) | 6449 (27 %) | 17262 (71 %) | 6449 (27 %) | 6449 (27 %) |
| 7 | 20636 (100 %) | 15361 (74 %) | 13899 (67 %) | 20636 (100 %) | 15361 (74 %) |
| 8 | 15434 (71 %) | 21615 (100 %) | 21615 (100 %) | 21615 (100 %) | 21615 (100 %) |
| 9 | 18238 (100 %) | 18238 (100 %) | | 18238 (100 %) | 18238 (100 %) |
| 11 | | 14419 (70 %) | | | 14419 (70 %) |
| 12 | | | | 19006 (96 %) | |
| 14 | 7902 (64 %) | 9764 (80 %) | 11900 (97 %) | 11900 (97 %) | 9764 (80 %) |
| 17 | 8703 (100 %) | 8703 (100 %) | | 8703 (100 %) | 8703 (100 %) |
| 19 | | 2965 (61 %) | 4286 (100 %) | | 2965 (61 %) |
| 20 | | 9852 (100 %) | 9852 (100 %) | 9852 (100 %) | 9852 (100 %) |
| 21 | | | 4879 (87 %) | | |

Table S8: **Assessing the stability of the algorithm, by rerunning the analysis on the Crohn's disease data. The list of large clusters identified by our method for each of the five runs.**
[a] The chromosome to which the SNPs in the group belong.
[b] The size of the SNP group is the number of SNPs that belong to the group. In parenthesis: size as percentage of total SNPs on the chromosome.

## S5 Software

The algorithm was implemented in the hierGWAS Bioconductor package, available for download from the Bioconductor website. The package allows to correct for confounding variables, as well as analyzing continuous phenotypes.

# References


Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag, Berlin.

Delaneau, O., Zagury, J., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*, **10**(1), 5–6.

Loh, P. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, **15**, 3065–3105.

Mandozzi, J. and Bühlmann, P. (2015). Hierarchical testing in the high-dimensional setting with correlated variables. *J Am Statist Assoc (published online DOI: 10.1080/01621459.2015.1007209)*.

Nagelkerke, N. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, **78**(3), 691–692.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ. Press.

Purcell, S. *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Int J Epidemiol*, **81**(3), 559575.